STAT 527

4/3/23

Lecture 3

- optimal h and Bias-Var for K.D
- KDE

LIII kernel, nn density est

- OH MMP
- Tue 4/4 at 4pm CSSS Conference Room PDL level LL

JP office h
Mon 1:30 - 2:30
CSSS Conf. Room

Lecture II – Nearest Neighbor and Kernel predictors

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

STAT/BIOST 527
Spring 2023

1. Nearest-Neighbor predictors ✓
2. Kernel predictors ✓
3. An elementary analysis of Kernel Regression ←
4. Bias, Variance and h for $x \in \mathbb{R}$ ←

h^* theoretical

in practice : CV next lecture

Regression ~~example~~ by N-W

A special case in wide use is the Nadaraya-Watson regressor

$$\hat{y}(x) = \frac{\sum_{i=1}^{n} b\left(\frac{||x-x^i||}{h}\right) y^i}{\sum_{i=1}^{n} b\left(\frac{||x-x^i||}{h}\right)}. = \sum w_i \, b_h(x, x^i) \quad (4)$$

In this regressor, $f(x)$ is always a convex combination of the y^i's, and the weigths are proportional to $b_h(x, x^i)$.
The Nadaraya-Watson regressor is biased if the density of P_X varies around x.

Assumed $y(x) = f(x) + \varepsilon$ — true !! — noise

$\hat{y}(x) - y(x) =$ error ← depends on ε

depends on $x^{1:n}, y^{1:n}$ data

Kernel regression by Nadaraya-Watson

$$\hat{y}(x) \;=\; \frac{\sum_{i=1}^{n} b\left(\frac{||x-x^i||}{h}\right) y^i}{\sum_{i=1}^{n} b\left(\frac{||x-x^i||}{h}\right)} \tag{6}$$

Let $w_i = \dfrac{b\left(\frac{||x-x^i||}{h}\right)}{\sum_{i'=1}^{n} b\left(\frac{||x-x^{i'}||}{h}\right)}$.

Assumptions

A0 For simplicity, in this analysis we assume $x \in \mathbb{R}$.

A1 There is a true smooth[1] function $f(x)$ so that

$$y \;=\; f(x) + \varepsilon, \tag{7}$$

where ε is sampled independently for each x from a distribution P_ε, with $E_{P_\varepsilon}[\varepsilon] = 0$, $Var_{P_\varepsilon}(\varepsilon) = \sigma^2$.

A2 The kernel $b(z)$ is smooth, $\int_{\mathbb{R}} b(z)dz = 1$, $\int_{\mathbb{R}} zb(z) = 0$, and we denote $\sigma_b^2 = \int_{\mathbb{R}} z^2 b(z)dz$, $\gamma_b^2 = \int_{\mathbb{R}} b^2(z)dz$.

In this first analysis, we consider that the values x, $x^{1:N}$ are fixed; hence, the randomness is only in $\varepsilon^{1:N}$.

[1] with continuous derivatives up to order 2

Expectation of $\hat{y}(x)$ – a simple analysis

Expanding f in Taylor series around x we obtain

$$f(x^i) = f(x) + f'(x)(x^i - x) + \frac{f''(x)}{2}(x^i - x)^2 + o((x^i - x)^2) \tag{8}$$

We also have

$$y^i = f(x^i) + \varepsilon^i. \tag{9}$$

We now write the expectation of $\hat{y}(x)$ from (6), replacing in it y^i and $f(x^i)$ as above. What we would like to happen is that this expectation equals $f(x)$. Let us see if this is the case.

$$E_{P_\varepsilon^n}[\hat{y}(x)] = E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i y^i\right] = E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i \left(f(x^i) + \varepsilon^i\right)\right] \tag{10}$$

$$= \sum_{i=1}^n w_i f(x) + \sum_{i=1}^n w_i f'(x)(x^i - x) + \sum_{i=1}^n w_i \frac{f''(x)}{2}(x^i - x)^2 + \underbrace{E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i \varepsilon^i\right]}_{=0} \tag{11}$$

$$= f(x) + \underbrace{f'(x)\sum_{i=1}^n w_i(x^i - x) + \frac{f''(x)}{2}\sum_{i=1}^n w_i(x^i - x)^2}_{\text{bias}} \tag{12}$$

In the above, the expressions in red depend of f and x, those in blue depend on $x^{1:n}$.

given $\mathcal{D} = (x^{1:n}, y^{1:n})$, x

$$E[\hat{y}(x) \mid \mathcal{D}, f, h, \text{noise } (0, \sigma^2)] = f(x) + \text{bias}_h(f', p_x') + \text{bias}_h(f'')$$

prediction

bias $E[\hat{y}(x) - f(x)] =$

Var $\hat{y}(x) \mid \mathcal{D}, f, h, \text{noise}$

- $\text{MSE}[\hat{y}(x)] = \text{bias}^2(\hat{y}(x)) + \text{Var}\,\hat{y}(x)$

Mean Squared Error

average over $\mathcal{D}$: $x^{1:n} \sim p_x$

$y = f(x) + \varepsilon$

- $\text{MSE}[\hat{y}(x) \mid p_x, f, h, n, \text{noise } \sigma^2]$
- $\text{MISE} = E_{p_x}[\text{MSE}\,\hat{y}(x)] = \int_{\mathbb{R}} \text{MSE}[\hat{y}(x)]\, p_x\, dx$

Integrated

Bias, Variance and h for $x \in \mathbb{R}$

2

The **bias** of $\hat{y}$ at x is defined as $E_{P_X^n} E_{P_\varepsilon^n} [\hat{y}(x) - f(x)]$.

$$E_{P_X^n} E_{P_\varepsilon^n} [\hat{y}(x) - f(x)] = h^2 \sigma_b^2 \left(\frac{f'(x) p_X'(x)}{p_X(x)} + \frac{f''(x)}{2} \right) + o(h^2) \tag{13}$$

The **variance** $\hat{y}$ at x is defined as $Var_{P_X^n P_\varepsilon^n}(\hat{y}(x))$.

$$Var_{P_X^n} P_\varepsilon^n(\hat{y}(x)) = \frac{\gamma^2}{nh} \sigma^2 + o\left(\frac{1}{nh}\right). \tag{14}$$

The **MSE (Mean Squared Error)** is defined as $E_{P_X^n} E_{P_\varepsilon^n} \left[(\hat{y}(x) - f(x))^2 \right]$, which equals

$$MSE(x) = \textbf{bias}^2 + \textbf{variance} = h^4 \sigma_b^4 \left(\frac{f'(x) p_X'(x)}{p_X(x)} + \frac{f''(x)}{2} \right)^2 + \frac{\gamma_b^2}{nh} \sigma^2 + \dots \tag{15}$$

- $MISE = \int MSE \, p_X \, dx =$

[2] After []

Optimal selection of h

$MISE_{(h)} = h^4 \blacksquare + \frac{1}{nh} \blacksquare$

If the MSE is integrated over $\mathbb{R}$ we obtain the **MISE**$= \int_{\mathbb{R}} MSE(x) p_X(x) dx$.
The kernel width h can be chosen to minimize the MISE, for fixed f, p_X and b.
We set to 0 the partial derivative

find min MISE

$$\frac{\partial MISE}{\partial h} = h^3 \left(\blacksquare \right) - \frac{(\blacksquare)}{nh^2} = 0. \quad (16)$$

It follows that $h^5 \propto \frac{1}{n}$, or

$n' = 100{,}000\, n$

$\Downarrow$

$h' = \frac{h}{10}$ (17)

$$h^* \propto \frac{1}{n^{1/5}}.$$

In d dimensions, the optimal h depends on the sample size n as

$$h^* \propto \frac{1}{n^{1/(d+4)}}. \Rightarrow n^{-\left(1-\frac{1}{d+4}\right)} = n^{-\frac{4}{5}} \quad (18)$$

slower

The MSE with optimal h decreases as $\sim \frac{1}{n} n^{1/(d+4)}$

Compare this with the MSE of the mean of a distribution, which decreases $\sim \frac{1}{n}$

Ex: $h \sim \frac{1}{n^\alpha}$ $\alpha > 0$

MSE rate?

$\bar{x} = \frac{1}{n} \sum x^i$ estimates $E_{P_X}[X]$

$MISE \sim \frac{1}{n}$

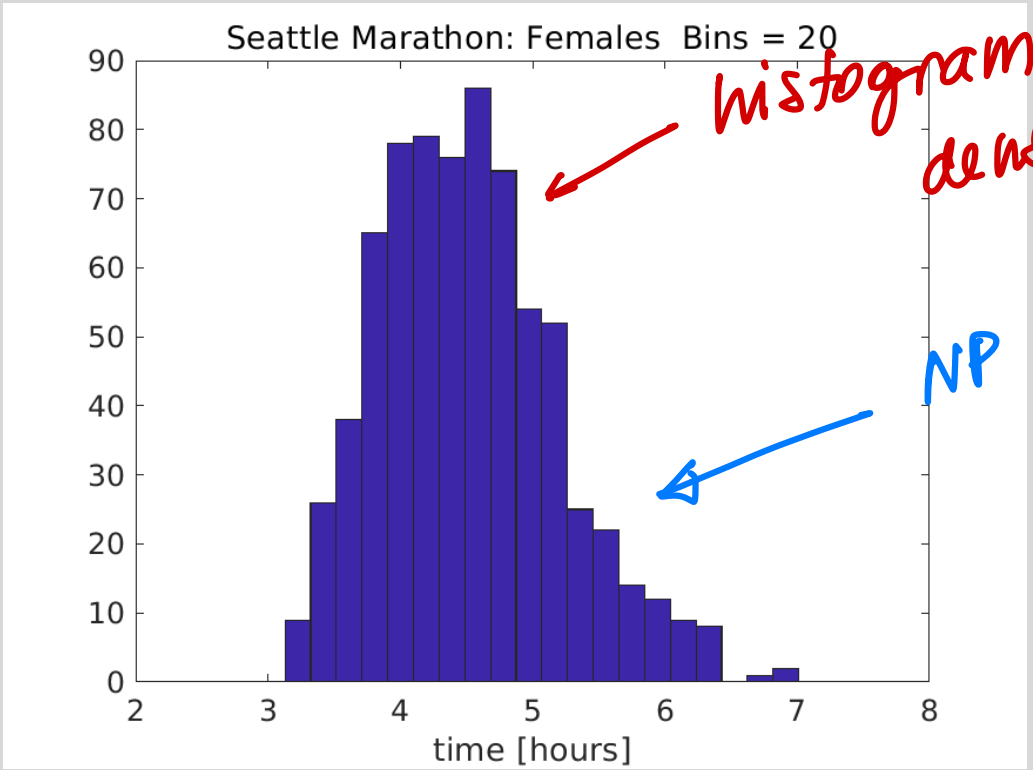
Lecture III – Kernel and k-NN density estimation

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

STAT/BIOST 527
Spring 2023

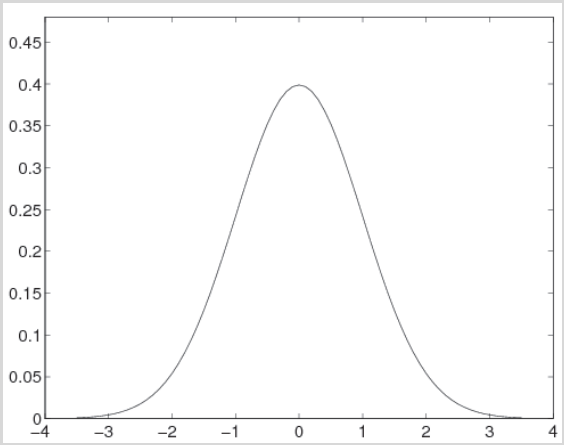
Arbitrary shaped densities

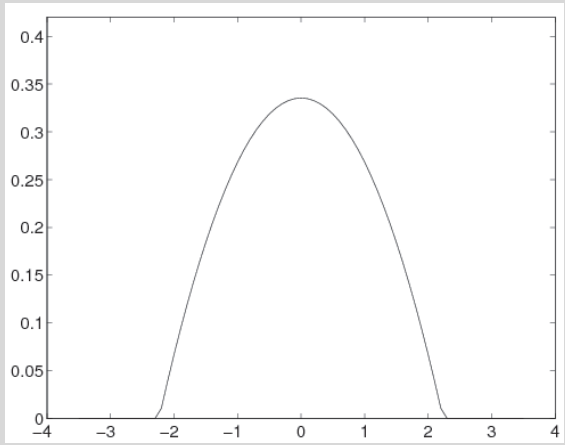


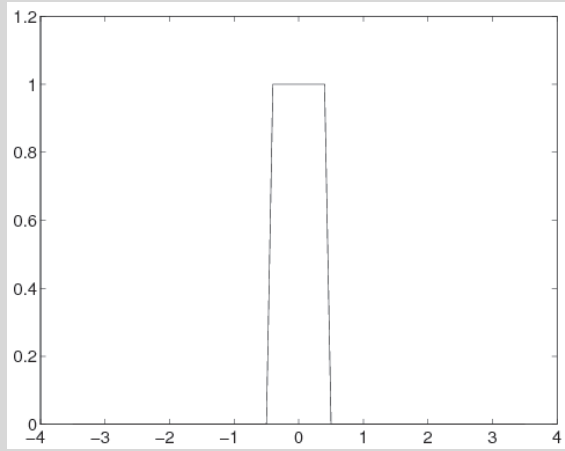
histogram is also NP density estimator

NP regression, density estimation = "smoothing"

Kernels







$E[b(z)] = 0$

$\int_{\mathbb{R}} b(z)\,dz = 1$. a density!

$b \geq 0$

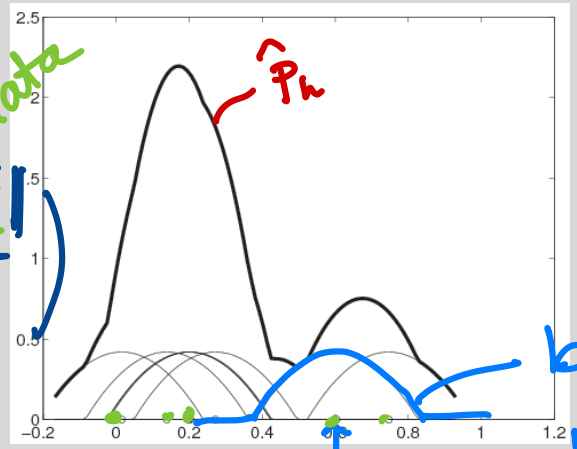
also, usually $\int_{\mathbb{R}} z^2 b(z) = 1$ ← different b's can use same h

Kernel density estimators

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} b_h(x, x^i)$$

$$= \frac{1}{nh}\sum_{i=1}^{n} b\left(\frac{\|x - x^i\|}{h}\right)$$

$$\mathcal{F}_{h,b} = \{\hat{p}_h\}$$



$$\frac{1}{h} b\left(\frac{\|x - x^i\|}{h}\right)$$

z

change of variable

$$\int_{-\infty}^{\infty} b(z)\,dz = 1$$

$$\tilde{z} = \frac{z}{h}$$

h = smoothness parameter

$h \uparrow$ $\hat{p}_x$ smoother

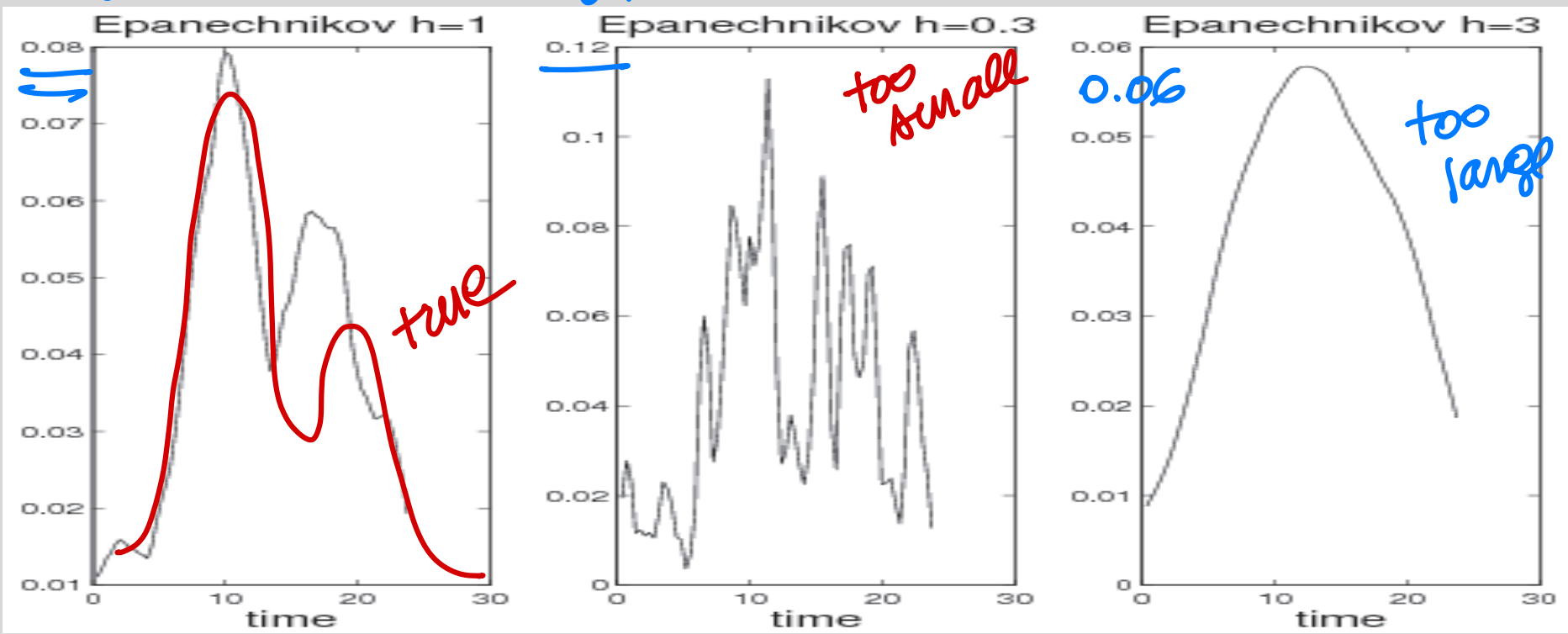
Var ↘

bias ↗

$$\int_{\mathbb{R}} b_h(z)\,dz = 1$$

for all $h >$

Choosing h


0.08
0.12
Epanechnikov h=1
Epanechnikov h=0.3
Epanechnikov h=3
true
too small
too large
0.06
time
time
time